Structural Similarity-based Link Prediction in Complex Networks

Thesis submitted in partial fulfillment

for the Award of Degree

DOCTOR OF PHILOSOPHY

by

AJAY KUMAR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY),

VARANASI-221 005

Roll No: 16071005

2021

Certificate

It is certified that the work contained in this thesis entitled “Structural Similarity-based Link Prediction in Complex Networks” by “AJAY KUMAR” has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy and SOTA for the award of Ph.D. Degree.

Dr. Bhaskar Biswas

Associate Professor

Department of Computer Science and Engineering

Indian Institute of Technology (Banaras Hindu University),

Varanasi-221 005

DECLARATION BY THE CANDIDATE

I, AJAY KUMAR, certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of Dr. Bhaskar Biswas from July 26, 2016 to October 5, 2021, at the Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and have not included them in this thesis and have not cited as my own work.

Ajay Kumar

Date :

Signature of the Student

(AJAY KUMAR)

Place : Varanasi

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

Signature of Supervisor

(Dr. Bhaskar Biswas)

Signature of Head of Department

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis : **Structural Similarity-based Link Prediction in Complex Networks**

Name of the Student : **AJAY KUMAR**

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the ”DOCTOR OF PHILOSOPHY”.

Ajay Kumar

Date :

Place : Varanasi

Signature of the Student

(AJAY KUMAR)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author’s personal use provided that the source and the Institute’s copyright notice are indicated.

This thesis is dedicated to my parents and family members
for their endless love, support and encouragement

PREFACE

Link prediction in complex networks (e.g., social networks, biological networks, citation networks, etc.) has attracted increasing attention from both physical and computer science communities. The algorithms can be used to extract missing information, identify spurious interactions, evaluate network evolving mechanisms, and so on. Its study is crucial to the analysis of the evolution of networks. Lots of works employing different types of methodologies of link prediction are available. Most of them are based on structural or topological properties as extracting these features are easy in computation. Though not all of them are efficient to extract. Most social networks exhibit some basic features like Small-world phenomenon, clustering and scale-free. Their corresponding measures are average path length, clustering coefficient and degree distribution respectively. In this thesis, these features are explored for calculating similarity measures of node-pairs in link prediction.

Many real-world networks show tendency of being organized in clusters that are quantified by clustering coefficient. This measure extracts local structural or topological information which are efficient to compute. The notion of mutual relationships, captured by common neighbors, are building blocks of many existing seminal works like Adamic-Adar index, resource allocation index, etc. The notion of common neighbors is further expanded to higher level. Based on clustering coefficients of level-2 common neighbors, a new algorithm CCLP2 is proposed to predict missing links in networks. CCLP2 extracts higher level clustering information of nodes which proved to be more informative and discriminating feature for link prediction as shown by the empirical results.

Exploring level-2 clustering information are useful discriminating feature but confined to neighbors of neighbors information. This might limit the prediction capability and hence, more local information are extracted using path feature. By employing higher order paths as discriminating features missing link are predicted in networks. The proposed method, called SHOPI, is based on resource allocation process in networks where the source node sends some resources as information to a destination node. The amount of information received by the destination derives the similarity score between them. Higher the information received by destination from the source represents higher similarity. SHOPI ensures to reach maximum information by restricting the information leaks through their common neighbor nodes. Empirical results on several networks validates the performance of SHOPI.

Acknowledgements

This thesis is the end of my journey in obtaining my Ph.D. I have not traveled in a vacuum in this journey. This thesis has been kept on track and been seen through to completion with the support and encouragement of numerous people including my well wishers, my friends, colleagues and various institutions. At the end of my thesis, I would like to thank all those people who made this thesis possible and an unforgettable experience for me. At the end of my thesis, it is a pleasant task to express my thanks to all those who contributed in many ways to the success of this study and made it an unforgettable experience for me.

Firstly, I would like to express my sincere gratitude to my advisor **Dr. Bhaskar Biswas** for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I specially thank the members of my Research Progress Evaluation Committee, **Dr. R. S. Singh** and **Dr. N. S. Rajput** for their invaluable suggestions regarding the thesis, their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

I would like to express my sincere thanks for **Prof. Rajiv Srivastava**, Head, Computer Science and Engineering Department, for his kindness and valuable support in carrying out the research. I express my sincere thanks to the faculty members, **Prof. K. K. Shukla**, **Prof. A. K. Tripathi**, **Prof. R. B. Mishra**, **Dr. Ravi Shankar Singh**, **Dr. Vinayak Srivastava**, **Dr. R. N. Choudary**, **Dr. H. P. Gupta**, **Dr. T. Dutta**, **Dr. A. K. Singh** and staffs of the department.

I am also thankful to all my friends for the fun moments in my PhD student life. I wish to express my deep sense of thanks to my seniors **Dr. Dharmendra Prasad Mahato**, **Dr. Harish Kumar Shakya**, **Dr. Anupam Biswas**, **Dr. Vinay Kumar**, **Dr. Ashish Kumar Maurya**, and **Dr. Vibhav Singh** for being a great friends and the best advisers I could ever have. Their advice, encouragement and critics were always the source of inspiration. This dissertation would not have been possible without their invaluable suggestions and persistent help. I wish to extend my gratitude towards all my lab mates **Mr. Shashank**

Sheshar Singh, **Mr. Shivansh Mishra**, **Ms. Sneha Mishra**, **Mr. S. P. Dwivedi**, **Mr. Naveen Mani Upadhyay**, and **Mr. Vishal Shrivastva** whose company helped me a lot. I am also thankful to **Mr. Yadul Raghav** and **Mr. Aman Gupta** for the collaborations and contributions to some of the projects in this dissertation. I am specially thanks to **Dr. Kuldeep Singh** without him this study might not come to an end.

Most importantly, my deepest gratitude is for my family for their constant support, inspiration, guidance, and sacrifices. My parents were constant source of motivation and inspiration. Their affection and guidance was instrumental in me choosing Engineering and eventually continuing on to my Ph.D. I would like to thank my mother, who was also my childhood teacher and is always there to stands by my side.

Last but not the least, I would like to special thanks is to my wife, elder brother **Mr. Sanjay Kumar** and younger brothers for supporting me unconditionally throughout my study. I sincerely thank all of them who contributed in for helping me to see the light at the end of every scary tunnel during my Ph.D.

Ajay Kumar

- AJAY KUMAR

Contents

List of Figures

List of Tables

Abbreviations

CCLP2	**Level-2** Node **C**lustering **C**oefficient-based **L**ink **P**rediction
SHOPI	Link Prediction in Complex Networks based on **S**ignificance of **H**igher-**O**rder **P**ath **I**ndex
AUROC	**A**rea **U**nder the **R**eceiver **O**perating **C**haracteristic Curve
AP	**A**verage **P**reccision
AUPR	**A**rea **U**nder the **P**recision-**R**ecall Curve
LCP	**L**ocal **C**ommunity **P**aradigm
LCL	**L**ocal **C**ommunity **L**inks
CAR	**C**annistraci-**A**lanis-**R**avasi
L3	Path of length 3
NSI	**N**eighbor **S**et **I**ndex
CCLP	**C**lustering **C**oefficient-based **L**ink **P**rediction
NMF	**N**on-negative **M**atrix **F**actorization
DCP	**D**egree related **C**lustering ability **P**ath
DR	**D**egree of **R**obustness

Symbols

$G(V,E)$	A social network with vertex set V and edge set E
A	Adjacency matrix of a network
n	The number of nodes in the network ($\|V\|$)
m	The number of Edges in the network ($\|E\|$)
$\Gamma(z)$	The neighbors set of node z
k_z	Degree of the node z
$t(z)$	Number of triangle passing through the node z
λ_1	Maximum eigen value of a matrix
β	dumping factor
$\langle K \rangle$	Average degree of the a network
$\langle D \rangle$	Average path length of a network
$\langle C \rangle$	Average clustering coefficient of a network
r	Coefficient of assortativity of a network
H	Degree of heterogeneity of a network
ρ	Network density
D	Diagonal matrix
$\mathscr{U}$ and $\mathscr{V}$ are left and right singular vectors	
L	Laplacian matrix
$C(z)$	Clustering coefficient of node z
CN^2	Level-2 common neighbors
CC(z)	Level-2 clustering coefficient a node z

α_c	Level of confidence
D_f	Degree of freedom
$S(x,y)$	Similarity score between the node x and the node y
ψ	Penalization factor for longer paths
l_{max}	Length of maximum path in the network
i_1, i_2	Intermediate nodes
$I_1 ....I_8$	Information flow in the network
I	Identity matrix